Industrial and Organizational Psychology
PERSPECTIVES ON SCIENCE AND PRACTICE

Industrial and Organizational Psychology, **7** (2014), 478–500.
Copyright © 2014 Society for Industrial and Organizational Psychology. 1754-9426/14

FOCAL ARTICLE

Corrections for Criterion Reliability in Validity Generalization: A False Prophet in a Land of Suspended Judgment

JAMES M. LEBRETON
Pennsylvania State University

KELLY T. SCHERER
Purdue University

LAWRENCE R. JAMES
Georgia Institute of Technology

Abstract

The results of meta-analytic (MA) and validity generalization (VG) studies continue to be impressive. In contrast to earlier findings that capped the variance accounted for in job performance at roughly 16%, many recent studies suggest that a single predictor variable can account for between 16 and 36% of the variance in some aspect of job performance. This article argues that this "enhancement" in variance accounted for is often attributable not to improvements in science but to a dumbing down of the standards for the values of statistics used in correction equations. With rare exceptions, applied researchers have suspended judgment about what is and is not an acceptable threshold for criterion reliability in their quest for higher validities. We demonstrate a statistical dysfunction that is a direct result of using low criterion reliabilities in corrections for attenuation. Corrections typically applied to a single predictor in a VG study are instead applied to multiple predictors. A multiple correlation analysis is then conducted on corrected validity coefficients. It is shown that the corrections often used in single predictor studies yield a squared multiple correlation that appears suspect. Basically, the multiple predictor study exposes the tenuous statistical foundation of using abjectly low criterion reliabilities in single predictor VG studies. Recommendations for restoring scientific integrity to the meta-analyses that permeate industrial–organizational (I–O) psychology are offered.

In the latter part of the 20th century, industrial–organizational (I–O) psychologists readily embraced a new statistical technology designed to guide quantitative reviews of applied psychology: meta-analysis. One particular flavor of meta-analysis was originally labeled *validity generalization* (VG; Schmidt & Hunter, 1977) to highlight its unique focus on ascertaining whether the evidence supporting the criterion-related validity of inferences drawn from scores on various employment tests would generalize across different situations. This form of meta-analysis has also been labeled psychometric meta-analysis (PMA) to emphasize that the meta-analytic procedure involves making "corrections" to observed correlation coefficients based on psychometric artifacts (e.g., measurement error and range restriction; Hunter & Schmidt, 2004).

Results of VG and PMA studies continue to be impressive. For example, general mental ability is said to have an average,

Correspondence concerning this article should be addressed to James M. LeBreton.
E-mail: james.lebreton@psu.edu
Address: Department of Psychology, Pennsylvania State University, 140 Moore Building, University Park, PA 16802

The authors would like to thank John Binning for his constructive comments on an earlier version of this article.

https://doi.org/10.1017/S1754942600006775 Published online by Cambridge University Press

corrected, criterion-related validity of .53 in predicting job performance (based on correcting for criterion unreliability and range restriction; Hunter & Hunter, 1984). Structured interviews may attain corrected validities in the .47 to .60 range in the prediction of job performance (Huffcutt & Arthur, 1994). Perceptual speed has an average corrected validity of .47 against clerical performance (Schmidt, 1992). Biographical data are believed to predict performance with mean corrected validities in the mid .30s (Rothstein, Schmidt, Erwin, Owens, & Sparks, 1990). Integrity tests have average corrected validities of .41 against job performance (application samples) and .47 against counterproductive behaviors (all samples; Ones, Viswesvaran, & Schmidt, 1993).

Many VG studies, taken individually, suggest that a single intellectual, cognitive, or personality trait can account for upwards of 16–36% of the variance in some aspect of job performance (or even more, if corrections are also made for predictor [trait] unreliability). The days when 16% of the criterion variance (or a validity of .40) was the maximum that could be explained by an individual test (Ghiselli & Brown, 1955) are gone, as are the days when validities in the .20s and .30s were commonplace in the reports of "well-done" validity studies.

The curious psychologist is left wondering what precipitated this boost in validities and variance explained? Perhaps better science is needed in the form of (a) improved measurement instruments and technologies, (b) more sophisticated sampling procedures, or (c) superior research designs? Well, not really. With a few notable exceptions, we are still relying on the same (largely self-report, survey-based) measurement procedures, the same small samples, and the same bivariate correlational validity studies. So, what has engendered this bountiful enhancement in validities?

A Culture of Corrections

We believe the boosts in criterion-related validity coefficients may be traced to the practice of correcting observed coefficients for attenuation due to unreliability in the criterion (and sometimes the predictor) and direct range restriction in the predictor. What changed then was not better science but the historical cautiousness in applying correction equations to validity coefficients. The conservativeness originally championed by Spearman (1910) has given way to a pervasive and largely uncensored use of corrections.

The genesis for this "culture of corrections" can be traced to desires to estimate relationships devoid of statistical artifacts. For example, latent variable procedures such as structural equation modeling (SEM) frame the opportunity to employ estimates of perfectly reliable variables in studies of covariation as a major advance in science (James, Mulaik, & Brett, 1982). No less dedicated to the pursuit of truth and scientific principle are the VG and PMA procedures (Schmidt, 1992). Such procedures are used to conduct meta-analyses on correlations corrected for various types of attenuation. It is the belief that corrected coefficients yield greater insight into scientific truths that has engendered the current culture of corrections. Researchers operating in this new culture are more prone to compute, and to be receptive to, corrected coefficients. Apparently, reviewers and journal editors are more likely to publish studies invoking such corrections.

To be clear, it is not our intent to place obstacles between scientists and the seeking of truth via corrected coefficients. It would be unfair to question corrections in VG/PMA analyses without also questioning their use in latent variable analyses such as SEM. On the contrary, we believe both latent variable analysis and meta-analysis, when properly used, represent useful tools in the toolbox of the applied psychologist. However, we do feel that it is reasonable for someone to inquire about the statistical values being used to make said corrections.

In addition, our concern is not with the "correctness" of correction equations nor is it with statistical procedures such as meta-analysis or SEM (both of which

https://doi.org/10.1017/S1754942600006775 Published online by Cambridge University Press

may make use of these equations). Instead, our concerns reside with the statistical values being used in these equations and the inferences drawn from the corrected correlation coefficients. We are specifically interested in the reliability values used in corrections for attenuation due to unreliability in criteria assessed via ratings of job performance.

Correcting for Unreliability in Performance Ratings

Viswesvaran, Ones, and Schmidt (1996) concluded that (a) job performance is typically assessed by ratings, (b) the reliability of ratings should be estimated via an interrater reliability analysis, and (c) the mean interrater reliability for supervisor ratings of job performance is approximately .52. Thus, if a study to be included in a meta-analysis fails to report criterion reliability, and the criterion is based on supervisor ratings, then the best estimate of the missing interrater reliability is .52. Consequently, this is the value that should be inserted into one of the various VG/PMA equations to estimate means and variances of estimated true correlations.

It is instructive to illustrate the consequence of using .52 as an estimate of reliability. Using the standard correction for attenuation, an observed criterion-related validity of .25 becomes a .35 (i.e., $.25/\sqrt{.52}$), .30 becomes a .42, .35 becomes a .49, and .40 becomes a .55. So simply by applying corrections for attenuation using a criterion reliability of .52, we obtain an 89% increase (i.e., $[.55^2 - .40^2]/.40^2 = .89$) in what is regarded as maximum likely efficiency for a single predictor.

How far are we willing to go with such logic? Suppose that we place a reliability of .16 in the correction equation, giving $.40/\sqrt{.16}$ The corrected correlation is estimated to be 1.0. But no one would do this, you say. Agreed; would a reliability of .26 be acceptable? No; well then would .36 be allowed? Still no; let's jump to .46. Now are we okay? This is very close to .52. Where do we draw the line? What are the minimally accepted standards for reliability?

Nunnally and Bernstein (1994) may be of assistance here. Regarding tests, these authors stated: "A reliability of .80 may not be nearly high enough in making decisions about individuals . . . If important decisions are being made with respect to specific test scores, a reliability of .90 is the bare minimum, and a reliability of .95 should be considered the desirable standard" (p. 265). More recently, Cascio and Aguinis (2005) suggested that "If a procedure is to be used to compare one individual with another, reliability should be above .90" (p. 145; see also the standards for reliability recapitulated in Table 1).[1]

What then is a desirable standard for ratings of job performance? If we desire a reliability of .90 or .95 for the test scores that help determine whether people get jobs, is it not reasonable to expect the same standard of reliability for the criteria used to compare individual employees to one another when determining which employees (a) keep their jobs, (b) are promoted to new jobs, or (c) receive merit-based raises or bonuses based on their performance? Apparently we are not alone in our belief that criteria should also be reliable:

> Relevancy [is] the first requirement for a criterion . . . Reliability is the second requirement of a criterion. (p. 746; Smith, 1976)
>
> if criteria are to be useful, they must be measureable in a consistent manner. (p. 167; Ployhart, Schneider, & Schmitt, 2006)
>
> criterion measures must be reliable. (p. 73; Latham & Wexley, 1981)

1. The interested reader is also directed to Lance, Butts, and Michels' (2006) review of reliability standards, wherein these authors noted that the .70 minimum cutoff for reliability, which is normally attributed to Nunnally (1978), is actually a methodological urban legend that misrepresents his recommendations; and, this legend has resulted in the widespread use of reliability standards that (in most instances) are too low.

https://doi.org/10.1017/S1754942600006775 Published online by Cambridge University Press

https://doi.org/10.1017/S1754942600006775 Published online by Cambridge University Press

Table 1. *A Return to Psychometrics Standards: Standards of Reliability for Measures Used to Draw Inferences About Individuals*

Source	Observations and recommendations
Guilford (1936)	"The reliability of ratings increase with the number of judges. According to Rugg (1921), one should use the pooled ratings of not less than three independent judges. Symonds (1931, p. 96) demands at least eight judges. According to Bradshaw (1930) with the average trait a reliability coefficient of .80 may be obtained with 5 to 22 judges; a reliability of .90 may be obtained with 21 to 106 judges." (p. 279)
Anastasi (1968)	"[desirable reliability coefficients] usually fall in the .80s or .90s." (p. 78)
Womer (1968)	"Clearly a reliability of .95, no matter what method is used [to estimate reliability], is a good one, even for interpreting individual test scores. Clearly a reliability of .65 is not satisfactory for individual scores . . . one can 'accept' any reliability above .80, while hoping for coefficients in the high .80's." (p. 41)
Guilford and Fruchter (1973)	"Following the leadership of T. L. Kelley there has been general agreement that to be sufficiently reliable for discriminating very accurately between individuals, a test should have a minimum reliability coefficient of at least .94. Some have been more liberal in this regard, allowing a minimum of .90." (p. 91) "the fact remains that in practice we expect reliability coefficients to be in [the range of] .70 to .98." (p. 92)
Weiner and Stewart (1984)	"If you were to use scores obtained on a test to make decisions about specific individuals, you would want (and need) a level of reliability higher (generally .85 and above) than if you rest use were to summarize a group's performance or to carry out a research project . . . where *no decisions will be made that affect individuals*, reliability coefficients in the .60s and .70s are acceptable." (p. 69; italics added)
Aiken (1988)	"If a test is used to determine whether the mean scores of two groups of people are significantly different, then a reliability coefficient as low as .65 may be satisfactory. If the *test is used to compare one examine with another*, a coefficient of at least .85 is necessary." (p. 101; italics added)
Murphy and Davidshofer (2005)	"reliability estimates around .70 are likely to be regarded as low . . . reliability estimates lower than .60 usually are thought to indicate unacceptably low levels of reliability." (p. 150)
LeBreton and Senter (2008)	"the minimum acceptable level of reliability for psychological measures in the early stages of development is .70 (Nunnally, 1978). *Higher levels may be required of measures . . . used in advanced field research and practice."* (p. 839; italics added)
Meriac, Hoffman, Woehr, and Fleisher (2008)	"Although we recognize that the .52 value reported by Viswesvaran et al. (1996) has often been used in meta-analytic studies, we concur with opponents of using this value as an estimate of the reliability of job performance ratings. Specifically, we believe this value represents and underestimate of performance rating reliability and thus results in an overestimate of the true correlations of various predictors with job performance." (p. 1046)

> Most texts in industrial psychology contain lengthy lists of requirements for criteria . . . [these lists] might be reduced to three requirements: reliability, validity, and practicality. (pp. 150–151; Landy, 1985)
>
> A second general quality of criteria that is most important is reliability. Great stress was put upon the importance of test reliability. *Criterion reliability is equally important*. (p. 65; Womer, 1968; italics added)
>
> Meaningful evidence of predictor-criterion relationships *is conditional on* a number features, including . . . the availability of a relevant and reliable criterion measure. (p. 159; AERA, APA, and NCME, 1999; italics added)
>
> A relevant, reliable, and uncontaminated criterion measure(s) *must be obtained* or developed. (p. 14; Society for Industrial and Organizational Psychology [SIOP], 2003; italics added)

Now it is true that the reliabilities for scores from tests used to make decisions about which employees get/keep jobs are not often in the .90s. Many, however, are in the .80s, and some are in the .70s. With the exception of performance ratings, none are in the .50s.

One notable historic exception consists of projective tests of personality (e.g., Rorschach Inkblot Test; Thematic Apperception Test); such tests have mostly disappeared from applied psychology journals, due in no small part to concerns related to poor interrater reliability and other psychometric limitations (James & LeBreton, 2012; Lilienfeld, Wood, & Garb, 2000).

If organizational psychologists are uncomfortable using projective tests as predictors because these tests lack interrater reliability, why then do we seem so ready to embrace supervisory ratings that appear to show equally problematic psychometric characteristics? Why are performance ratings allowed to survive in spite of what most would agree is abjectly problematic measurement? Why do we allow observed validities to be corrected for unreliability in what appear to be fundamentally flawed variables and then act as if these corrected validities actually convey some sort of credible scientific information? Does anyone really believe that it makes sense to talk about a "perfectly reliable criterion" when the criterion begins with an observed interrater reliability of .52 for supervisor ratings (or .42 for peer ratings; Viswesvaran et al., 1996)? How exactly does a variable in which ~50% of the observed variance is measurement error variance become perfectly reliable?

It would seem that researchers would have instituted the necessary improvements, given that problems with performance ratings were documented 60 years ago in Guilford's (1954) classic text in psychometrics. Have not a large number of articles and books been written on the biases and errors that affect performance ratings, especially after the classic articles on these subjects by Feldman (1981) and Landy and Farr (1980)? Various problems with criteria in general, and performance ratings in particular, have been documented for nearly 100 years (cf. Aguinis, 2013; Austin & Villanova, 1992; Binning & Barrett, 1989; Cardy & Dobbins, 1994; Cooper, 1981; James, 1973; Murphy & Balzer, 1989; Murphy & Cleveland, 1995; Saal, Downey, & Lahey, 1980; Thorndike, 1920). Thus, we know what at least some of the problems are. If they are tractable, then why have they not been fixed?

Is it because these problems are intractable? Is it not possible to build ratings that can achieve high interrater reliabilities? If this is so, or even plausible (which history suggests that it is), then should we not experience some embarrassment when we insert .52 in corrections for attenuation because we suspect that the "theoretically perfect reliability" is not going to be even remotely approximated?

The value of .52 may or may not be an accurate estimate of interrater reliability. This issue is currently being debated elsewhere and has implications for the estimates used in VG and SEM studies (cf.

https://doi.org/10.1017/S1754942600006775 Published online by Cambridge University Press

DeShon, 2003; LeBreton, Burgess, Kaiser, Atchley, & James, 2003; Murphy & DeShon, 2000a, 2000b; Putka, Le, McCloy, & Diaz, 2008; Schmidt, Viswesvaran, & Ones, 2000; Viswesvaran et al., 1996). At issue here is the willingness of authors to use and leading journals to publish corrections for attenuation based on a criterion reliability of .52. Obviously, both authors and editors assume that .52 is a legitimate and accurate estimate of interrater reliability. If this estimate is later shown to be inaccurate or ill founded, then the plot thickens. However, for now, let us assume that the estimate of .52 is legitimate and accurate. We may then deal with the issue of whether substantive scientific judgments should be based on corrections that employ a ridiculously low reliability for a criterion to yield enhanced, sometimes much enhanced, estimates of corrected validity.

A Statistical Experiment

We have attempted to build an argument that we cannot base scientific conclusions on large corrections engendered by poor measurement. To extend this argument a bit beyond the discussion above, allow us to demonstrate some unanticipated consequences of inserting low reliabilities into correction equations. This demonstration begins by asking the question: What happens if we include multiple predictors in a selection study and conduct a multiple correlation analysis, applying corrections that are normally observed in a single predictor study?

Specifics are easily attained with illustrative data sampled from a series of meta-analyses in the area of employee selection (cf. Hermelin & Robertson, 2001; Schmidt & Hunter, 1998). For this illustration, we opted to examine the additive predictive power of the constructs assessed using general mental ability tests, integrity tests, structured employment interviews, and work sample tests. The upper portion of Table 2 contains the uncorrected, meta-analytic estimates linking predictors to performance. The middle portion of Table 2 contains the partially corrected correlations obtained after correcting for range restriction and criterion unreliability. The bottom portion of Table 2 contains the fully corrected correlations obtained after correcting for range restriction, criterion unreliability, and predictor unreliability.

Corrections for statistical artifacts were applied consistent with what Hunter and Schmidt (2004) would call a *full artifact distribution meta-analysis*. This approach to meta-analysis is typically used when information on artifacts is only sporadically available across primary studies.

Corrections for range restriction were based on *U* values (ratio of the restricted standard deviation to unrestricted standard deviation) of .66 for general mental ability (Hermelin & Robertson, 2001; Hunter, 1983 as cited in Hermelin & Robertson, 2001), .81 for integrity (Hermelin & Robertson, 2001; Ones et al., 1993), .65 for structured interviews (Hermelin & Robertson, 2001), and 1.00 for work samples. Constraining the final ratio to unity is consistent with the presentation of Schmidt and Hunter (1998); (Table 1, Footnote b) and provides a conservative test of our thesis.

Corrections for attenuation in the predictors were based on reliabilities of .90 for general mental ability tests (Schmidt & Hunter, 1998), .81 for integrity tests (Ones et al., 1993), .93 for structured interviews (Campion, Campion, & Hudson, 1994), and .77 for work sample tests (Schmidt, Hunter, & Outerbridge, 1986). Corrections for attenuation in supervisory ratings of job performance were based on an interrater reliability of .52. Corrections to the predictor intercorrelations, already corrected for range restriction, were based on the standard correction for attenuation in the predictor and the criterion.

After correcting for criterion unreliability and range restriction, we see that the "corrected" validities now run from .41 to .67. The .67 is for structured interviews and represents a 2.88 ($[.67^2 - .34^2]/ .34^2 = 2.88$), or 288%, proportional increase in the explanatory power of structured interviews.

https://doi.org/10.1017/S1754942600006775 Published online by Cambridge University Press

Table 2. *Uncorrected Correlation Matrix (a), Correlation Matrix Corrected for Range Restriction and Criterion Unreliability (b), and Fully Corrected Correlation Matrix (c)*

Variable	1	2	3	4	5
(a)					
1. Job performance	1.00				
2. General mental ability[a]	0.27	1.00			
3. Integrity[b]	0.25	0.00	1.00		
4. Structured interview[c]	0.34	0.23	0.20	1.00	
5. Work sample[d]	0.42	0.33	0.00	0.25	1.00
(b)					
1. Job performance	1.00				
2. General mental ability	0.54	1.00			
3. Integrity	0.41	0.00	1.00		
4. Structured interview[e]	0.67	0.30	0.20	1.00	
5. Work sample	0.58	0.33	0.00	0.25	1.00
(c)					
1. Job performance	1.00				
2. General mental ability	0.57	1.00			
3. Integrity	0.47	0.00	1.00		
4. Structured interview	0.70	0.33	0.23	1.00	
5. Work sample	0.66	0.40	0.00	0.30	1.00

[a]The uncorrected correlation between job performance and general mental ability was set to .27, which corresponds to the correlation obtained for moderately complex jobs (Table 3; Hermelin & Robertson, 2001).
[b]The uncorrected correlation between integrity and job performance was set to .25, which corresponds to the correlation obtained using predictive validity studies conducted with job applicants (Table 8; Ones et al., 1993) and corresponds to the samples used to generate the corrected correlation discussed in Schmidt and Hunter (1998). The correlation between integrity and general mental ability was set to zero per the recommendations of Schmidt and Hunter (1998), as was the correlation between integrity and scores on work sample. We were unable to locate meta-analytic estimates of the correlation between integrity tests and structured interviews; however, we set this value equal to .20, thus permitting for some overlap in the constructs measured by these two selection tools.
[c]The uncorrected correlation between structured interviews and job performance was set to .34, which corresponds to the correlation using highly structured interviews (p. 188, Huffcutt & Arthur, 1994; Table 3, Hermelin & Robertson, 2001). The correlation between (highly) structured interviews and general mental ability was set to .23 (Table 3, Huffcutt, Roth, & McDaniel, 1996). We were unable to locate meta-analytic estimates of the correlation between work sample tests and structured interviews; however, we set this value equal to .25 to allow for the possibility that some of the KSAs measured during a work sample might also be sampled during a structured interview.
[d]The average, uncorrected correlations between work sample tests and job performance were not presented in Schmidt and Hunter (1998) or Hunter and Hunter (1984). Consequently, following Hermelin and Robertson (2001), we estimated the uncorrected value by attenuating the corrected coefficients. Specifically we attenuated the corrected value of .54 (Table 10, Hunter & Hunter, 1984; Table 1, Schmidt & Hunter, 1998) for measurement error using a reliability of .60 (p. 79; Hunter & Hunter, 1984). Consequently, we set the uncorrected correlation between work sample tests and job performance to .42; the uncorrected correlation between work sample tests and general mental ability was set to .33 (Table 2; Schmidt et al., 1986);
[e]The correlation between highly structured interviews and general mental ability corrected for range restriction in interviews was set at .30 (Table 3; Huffcutt et al., 1996).

Next we conduct a multiple correlation analysis using the uncorrected data in Table 2 and obtain an R^2 of .289, suggesting that these four predictor variables account for 29% of the variance in supervisory ratings of job performance.

We then conduct a multiple regression analysis using the correlations corrected for range restriction and criterion unreliability and obtain an R^2 of .793, suggesting that 79% of the variance in job performance is theoretically accounted for by

https://doi.org/10.1017/S1754942600006775 Published online by Cambridge University Press

a weighted linear combination of general mental ability, integrity, the variables measured in structured interviews, and work sample tests. This represents a 1.74 ([.793-.289]/.289), or 174%, proportional increase in the multivariate explanatory power of our test battery.

Finally, we conduct the same multiple regression analysis using the fully corrected correlations and obtain an R^2 of .893, suggesting that 89% of the variance in the latent construct of job performance is explained by the latent constructs measured by our battery of four employment tests. The correction for predictor unreliability yields an additional 0.126 ([.893-.793]/.793), or 13%, proportional increase in the multivariate predictive validity of our test battery. So, with a sampling of constructs measured using only four selection tests, we are able to explain nearly 90% of the variance in the job performance construct.

A confirmed trait theorist might conclude, "we have long told you it's all about traits." Those with interactional or perhaps situational proclivities are probably a little uncomfortable. For, now only 10% of the variance in our job performance is available to be uniquely predicted by situational variables, additional individual difference variables, or interactions between the two.

To be more explicit, if the multiple correlation model described above is even remotely correct, then variables unlikely to make substantial, unique contributions to the understanding of job performance include:

1. additional individual difference variables (e.g., achievement motivation, proactive personality, core self-evaluations, self-esteem, general self-efficacy, trait affect, Big Five personality traits; aggression; narcissism, Machiavellianism), as well as various situational and perceptual variables such as . . .
2. work motivation (e.g., goal setting, goal orientation, self-regulation, resource allocation),
3. group or team inputs or processes (e.g., team cohesion, collective efficacy, team conflict),
4. type of work or task performed (e.g., autonomous vs. pooled vs. sequential work activities),
5. organizational climate (e.g., leader trust and support, work group warmth, role ambiguity, role conflict)
6. organizational culture (e.g., shared norms and standards for behavior),
7. job attitudes (e.g., job satisfaction, organizational commitment),
8. leadership and power,
9. equity, fairness, and justice,
10. work stress and strain,
11. job characteristics and job redesign,
12. mentoring,
13. organizational support,
14. training,
15. newcomer socialization,
16. compensation,
17. performance feedback and coaching,
18. human–factors/person–machine interface,
19. work–family conflict/work–family enrichment,
20. any number of other possible organizational interventions and/or random shocks to the system, or
21. interactions among the various combinations of the above.

In short, only 10% of the variance in job performance remains to be uniquely explained by the main, interactive, and polynomial effects engendered by the myriad specific variables nested within the 20 general classes of antecedent variables noted above. As scientists and practitioners, we must ask ourselves if this seems reasonable. What might the extant literature have to say about the relative impact of these 20 classes of antecedents?

Antecedents of Job Performance: Illustrative Examples

Table 3 provides a few illustrative references to meta-analyses summarizing

https://doi.org/10.1017/S1754942600006775 Published online by Cambridge University Press

Table 3. *Illustrative Examples of the Correlations Between Measures of Job Performance and Potential Antecedent Variables*

Study	Antecedent	r	$\hat{\rho}$
Fried et al. (2008)[a]	Role ambiguity	−0.15	−0.24
	Role conflict	−0.10	−0.15
	Job satisfaction	0.16	0.25
	Propensity to leave	−0.12	−0.19
Bowling, Eschleman, Wang, Kirkendall, and Alarcon (2010)[a,h]	Organization-based self-esteem	0.23	0.34
	General self-esteem	0.12	0.18
Tornau and Frese (2013)[c,d]	Personal initiative		0.22
	Proactive personality		0.15
	Taking charge		0.34
	Voice		0.30
Erdogan et al. (2012)[e]	Life satisfaction	0.14	
Riggle, Edmondson, and Hansen (2009)[f]	Perceived organizational support	0.16	0.18
Thomas, Whitman, and Viswesvaran (2010)[a,g,h]	Proactive personality	0.27	0.38
	Personal initiative	0.25	0.35
	Voice	0.40	0.59
	Taking charge	0.33	0.46
Hoobler, Hu, and Wilson (2010)[b,h]	Work-to-family conflict	−0.16	−0.19
	Family-to-work conflict	−0.12	−0.16
Ford, Cerasoli, Higgins, and Decesare (2011)[b,h]	Psychological well-being	0.33	0.42
	Depression	−0.15	−0.18
	Anxiety	−0.19	−0.24
Chiaburu and Harrison (2008)[d,h]	Coworker support	0.20	0.24
	Coworker antagonism	−0.10	−0.11
Joseph and Newman (2010)[d,h]	Emotional intelligence		
	Self-report mixed	0.32	0.47
	Self-report ability	0.17	0.23
	Performance based	0.16	0.18
Cooper-Hakim and Viswesvaran (2005)[b,h]	Work ethic	0.38	0.47
	Job involvement	0.14	0.18
	Affective organizational	0.22	0.27
Bhaskar-Shrinivas, Harrison, Shaffer, and Luk (2005)[d,h]	Cultural adjustment	0.13	0.16
	Interaction adjustment	0.16	0.18
	Work adjustment	0.30	0.37
Fried (1991)[b,h]	Job variety	0.15	0.21
	Job identity	0.12	0.17
	Job autonomy	0.15	0.21
	Job feedback	0.13	0.19
Salas, Rozell, Mullen, and Driskell (1999)[e]	Team building	0.14	
Rockstuhl, Dulebohn, Ang, and Shore (2012)[b,h]	LMX in cultures characterized as horizontal individualism	0.29	0.30
	LMX in cultures characterized as vertical collectivism	0.26	0.29

https://doi.org/10.1017/S1754942600006775 Published online by Cambridge University Press

Table 3. *Continued*

Study	Antecedent	r	$\hat{\rho}$
Barrick and Mount (1991)[a,h]	Extraversion	0.08	0.14
	Emotional stability	0.05	0.09
	Agreeableness	0.05	0.09
	Conscientiousness	0.15	0.26
	Openness to experience	0.02	0.04
Colquitt, Scott, and LePine (2007)[d]	Trust	0.26	0.33
Colquitt et al. (2001)[d]	Procedural justice	0.30	0.36
	Distributive justice	0.13	0.15
Meriac et al. (2008)[b]	Assessment center dimensions:		
	Communication		0.27
	Organizing and planning		0.35
	Problem solving		0.32
	Stress tolerance		0.18
Oh, Wang, and Mount (2011)[h,i]	Observer ratings of:		
	Conscientiousness	0.28	0.33
	Emotional stability	0.15	0.18
	Agreeableness	0.21	0.26
	Extraversion	0.19	0.24
	Openness/intellect	0.18	0.23

Note. When multiple measures of job performance were included in the meta-analysis, we selected those relying on supervisory ratings and/or relying on subjective measures of performance.
[a]Uncorrected correlations were corrected using an interrater reliability estimate of 0.52 for supervisory ratings.
[b]Uncorrected correlations were corrected using an internal consistency reliability estimate or some other estimate not based on the interrater reliability estimate of 0.52.
[c]Uncorrected correlations were not reported in the original article.
[d]Details concerning the exact values used to correct for measurement error in the criterion were not reported in the original article. In some instances this information was completely omitted, in other instances the authors simply noted that they relied on sample estimates of reliability and/or imputed average reliabilities using the data in their study.
[e]Corrected correlations were not reported in the original article.
[f]Original paper reported correct coefficients and average reliabilities for predictor (.90) and criterion (.87). These values were used to provide a rough estimate of the uncorrected correlation.
[g]Uncorrected correlations were corrected an interrater reliability estimate of 0.42 for peer ratings.
[h]The authors appropriately denoted the corrected coefficients using notation such as "estimated ρ"or "$\hat{\rho}$."
[i]Uncorrected correlations were corrected using an internal consistency reliability estimate as well as interrater reliability (.52) adjusted for the number of raters (i.e., not the average interrater reliability estimate of .52 for a single rater).

the predictive validity of a sampling of potential antecedents of job performance. A few things are worth noting from this table. First, the examples in this table are circumscribed—we have not exhausted all possible antecedents of performance. However, we do believe this table provides a general sense of the magnitude of relationships one might expect to see between a sampling of dispositional and situational variables. Second, if the multiple regression model described above is correct, then only 10% of the variance in the job performance construct remains to be uniquely explained by (a) the variables contained in Table 3 (and those not contained in Table 3, but derived from the 20 classes of antecedents described above), (b) all interactions among all variables in Tables 2 and 3, and (c) all higher-order polynomial effects for the variables in Tables 2 and 3.

Let us consider a further subset of variables drawn from Table 3. Specifically,

https://doi.org/10.1017/S1754942600006775 Published online by Cambridge University Press

Table 4. *Uncorrected Correlations for a Subset of Variables From Table 3.*

Variable	1	2	3	4	5	6	7	8	9
1. Job performance	1.00								
2. Work to family conflict	−0.16	1.00							
3. Family to work conflict	−0.12	0.39	1.00						
4. (Work) social support	0.20	−0.23	−0.09	1.00					
5. Job involvement	0.14	0.08	0.01	0.15	1.00				
6. (Work) role conflict	−0.10	0.36	0.20	−0.36	0.00	1.00			
7. (Work) role ambiguity	−0.15	0.24	0.19	−0.24	−0.07	0.49	1.00		
8. Job satisfaction	0.16	−0.25	−0.15	0.43	0.30	−0.41	−0.45	1.00	
9. Life satisfaction	0.14	−0.30	−0.20	0.32	−0.02	−0.31	−0.26	0.44	1.00

Note. Correlations with job performance and work-to-family conflict, $r = -.16$ (Hoobler et al., 2010); family-to-work conflict, $r = -.12$ (Hoobler et al., 2010); social support $r = .20$, (Chiaburu & Harrison, 2008); job involvement, $r = .14$ (Cooper-Hakim & Viswesvaran, 2005); work role conflict, $r = -.10$ (Fried et al., 2008); work role ambiguity, $r = -.15$ (Fried et al., 2008); job satisfaction, $r = .16$ (Fried et al.); and, life satisfaction, $r = .14$ (Erdogan et al., 2012). Uncorrected meta-analytic estimates of the intercorrelations between variables were obtained from Michel, Mitchelson, Kotrba, LeBreton, and Baltes (2009; Table 2, p. 206).

we will use the uncorrected correlations between supervisory ratings of job performance and measures of: work-to-family conflict ($r = -.16$; Hoobler, Hu, & Wilson, 2010), family-to-work conflict ($r = -.12$; Hoobler et al., 2010), social support ($r = .20$; Chiaburu & Harrison, 2008), job involvement ($r = .14$; Cooper-Hakim & Viswesvaran, 2005), work role conflict ($r = -.10$; Fried, Shirom, Gilboa, & Cooper, 2008), work role ambiguity ($r = -.15$; Fried et al., 2008), job satisfaction ($r = .16$; Fried et al., 2008), and life satisfaction ($r = .14$; Erdogan, Bauer, Truxillo, & Mansfield, 2012). Uncorrected meta-analytic estimates of the intercorrelations between these variables were obtained from Michel, Mitchelson, Kotrba, LeBreton, and Baltes (2009; Table 2, p. 206). An examination of this correlation matrix (see Table 4) yields the following insights:

1. The R^2 obtained by regressing supervisory ratings of job performance onto these variables is .08.
2. The average (absolute value) of the uncorrected criterion-related validity correlations in Table 4 is .15 (which is just slightly lower than the value of .18 obtained for all of the variables in Table 3).
3. The average (absolute value) of the uncorrected variable intercorrelations in Table 4 is .25.

Thus, a weighted linear combination of the eight variables in Table 4 explains roughly 8% of the variance in job performance. This isn't quite 10%, but it represents a sizeable portion of the variance remaining in job performance after accounting for the traits measured by our four selection tests.

What would happen if we extrapolated the information obtained for these eight variables to the 48 variables presented in Table 3? To simplify the analysis, we constrained all validities to be positive and set all predictor intercorrelations to +0.25. We then regressed performance on this rather large set of variables and obtained an R^2 of .60. Suddenly, the tenability of our "corrected" regression model with only four selection tests appears suspect. If our R^2 based on corrected validities is correct, then only 11% of the variance in job performance remains to be uniquely predicted by the situational and perceptual variables in Table 3. Thus, our corrected model can only be "correct" if the situational/perceptual variables in Table 3 are highly redundant with the variables measured in our selection test battery. Does

https://doi.org/10.1017/S1754942600006775 Published online by Cambridge University Press

this seem likely? Would we expect the constructs measured in our selection test battery (i.e., critical, job-specific KSAs) to be highly redundant with those constructs assessed by measures of organization-based self-esteem, work role ambiguity, role conflict, (work) social support, coworker antagonism, work-to-family conflict, work group warmth, job feedback, perceptions of leadership, trust, or procedural justice?

Let us consider what might happen when we integrate variables from Table 2 with those from Table 3. Take for example, the research of Colquitt, Conlon, Wesson, Porter, and Ng (2001) who reported an uncorrected bivariate correlation between perceptions of procedural justice fairness and job performance of .30 (see their Table 5, p. 436). Let us explore what happens if we add perceptions of procedural justice to the prediction model containing our four selection tests. We will work with the fully corrected correlation matrix presented at the bottom of Table 2 and simply add the uncorrected bivariate criterion related validity of .30 for job performance (Note: For our purposes, it is not necessary to correct this observed validity for measurement error in the predictor or criterion. As illustrated below, doing so would be statistical overkill.)

The theoretical intercorrelations between procedural justice and the four selection variables are set to zero. There is no reason that intelligent people should be more or less likely to perceive their organizations as just. There is also no reason to expect that correlations between a job applicant's behavior during a screening interview (or in a work sample) would be correlated with his or her subsequent subjective perceptions of procedural justice fairness once employed. The only correlation one might expect would be between integrity and procedural justice, due to the typical use of self-reports to measure both constructs. Inasmuch as we are addressing theoretically substantive relationships here, we excised any relationship due to method variance. Work samples and interviews might pick up on something that would later affect job attitudes, but there is nothing that is clearly indicated. Finally, and more importantly, procedural justice is simply serving as a surrogate for a great many variables described earlier under the 20 classes of job performance antecedents. Thus, by setting the intercorrelations to zero, we are merely hastening the inevitable of what is going to happen by placing any number of unmeasured variables (see Tables 3 and 4) in the equation alongside our original four selection tests.

So, what happens when we add the (uncorrected) correlation between procedural justice and job performance to the (corrected) correlations between performance and our four selection tests? Now we are able to explain 98.3% of the variance in job performance. Of course, if we "corrected" the observed correlation with procedural justice for measurement error, then the explained variance in job performance would have exceeded unity, which is obviously impossible. We could of course increase the degree to which we account for more than 100% of the variance by adding additional variables from Tables 3 or 4 to the equation; however, doing so would be overkill.

It is instructive to view what would have happened if we had used reasonable values for criterion reliability set by the psychometric standards we discussed earlier (and summarized in Table 1). If everything remained unchanged except that .90 was used as the estimate of criterion reliability, the R^2 based on the four selection variables plus procedural justice is .604.

So, with reasonable values for criterion reliability set by accepted scientific standards in psychometrics, corrected coefficients provide R^2s that (while large in magnitude) remain in the proper range (i.e., do not approach or exceed unity). However, when scientific standards are suspended and questionable values for reliability are used in the correction process, the R^2 values may easily wander off into improper territory. We do not see this in bivariate studies because validity would have to be larger than the square root of reliability in order to

https://doi.org/10.1017/S1754942600006775 Published online by Cambridge University Press

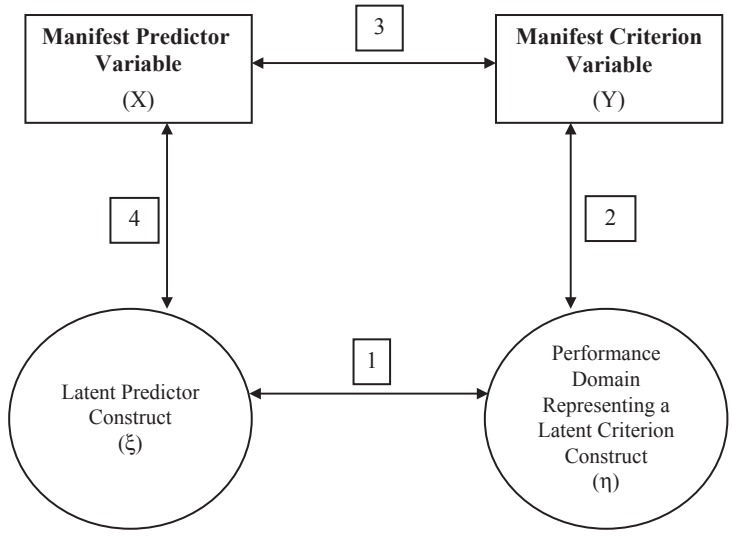


Figure 1. Basic construct validation model.

have an r^2 greater than 1.0. This would raise a red flag and to our knowledge has not yet been tried. But as a discipline, applied psychology has failed to realize that we have started down a slippery slope that produces a less dramatic but nonetheless specious result. Once we begin to suspend judgment about acceptable thresholds for criterion reliability (i.e., ignore nearly 100 years of advice from leading psychometricians) and allow a value as low as .52 into correction equations (or .42 for that matter; see Thomas, Whitman, & Viswesvaran, 2010), we step onto the slope. The multiple correlation analysis performed here picked up on the slippery slope by producing a grossly inflated R^2. It follows that the corrected bivariate components that engendered this inflated value possess a tenuous foundation.

Meaning and Interpretation of Corrected Coefficients

In addition to concerns with the values often used in correction equations, we also think it is important to ask whether psychologists are appropriately reporting and interpreting these corrected (or partially corrected) coefficients. That is, what information do corrected correlations confer, and how should they be reported in the literature?

If we define validity as the extent to which accumulated evidence supports inferences drawn from psychological measurements (cf. Binning & Barrett, 1989; Landy, 1986; Messick, 1995; SIOP, 2003), then the goals of the psychologist working on validation efforts should be to (a) accumulate evidence, (b) ensure that the evidence is of the highest quality possible, and (c) draw proper inferences from that evidence.

Consider the basic validation model presented in Figure 1 (Binning & Barrett, 1989). The two circles correspond to the latent (unobserved) constructs while the two squares correspond to the manifest (observed) indicators of these constructs. Four inferences link these two variables and two constructs. James (1973) labeled Inference 1 as a *constitutive* definition linking the latent constructs to one another and Inferences 2 and 4 the *epistemic* definitions linking the latent constructs to their observed variables. Within the context of I–O psychology, Inference 3 is often framed as the uncorrected criterion-related validity coefficient between some predictor (*X*) and some criterion (*Y*). James (1973) uses the term "P-plane" to denote the relationships among perceptual variables (i.e., manifest

https://doi.org/10.1017/S1754942600006775 Published online by Cambridge University Press

indicators; Inference 3) and "C-plane" to denote the relationships among latent constructs (Inference 1; see also Binning & Barrett, 1989; Margenau, 1950).

Scientific Interpretation

From a scientific perspective, the corrected coefficients represent a *theoretical approximation* of the relationship one might expect between the predictor and the criterion, assuming the criterion was "made perfectly reliable by lengthening . . . the criterion measure indefinitely" (p. 102, Gulliksen, 1950). From the perspective of the scientist, both predictors and criteria should be corrected for measurement error so that approximations of the C-plane relationships are as accurate as possible. Such are the corrections that are regularly made as part of a latent variable SEM analysis. The interpretation of such a corrected coefficient is that it is the *theoretical approximation* of the C-plane relationship obtained if one had access to an *infinitely long predictor* and an *infinitely long criterion* (i.e., perfectly reliable measures reflecting a one-to-one correspondence between the manifest variable and the latent constructs of interest).

As applied psychologists, we must ask ourselves how often we will encounter situations where data are available on an infinite number (or at least a very large number) of psychometrically parallel supervisors (cf. DeShon, 2003; Murphy & DeShon, 2000a; Putka et al., 2008)? For, if such data do not exist, then the interpretation of a corrected coefficient when the criterion is measured using supervisory ratings has little scientific (or practical) value.

Practical Interpretation

"Not so" claims the advocate of VG. To him or her, this is not a theoretical approximation but something called an "operational validity." The logic underlying the computation and interpretation of the operational validity is that one should make corrections for criterion unreliability (but not predictor unreliability) and then interpret the resulting (partially corrected) coefficient as what might be achieved using the predictor in actual application or operation in a selection context. That is, the operational validity is mathematically a hybrid coefficient for which the observed correlation is corrected for measurement error in one measure but not the other. Yet, the inference that is often drawn seems suggestive of how we should expect a selection test to perform "operationally" or "in practice." Such an inference is conceptually muddied; it confounds estimates of C-plane relationships with a P-plane inference.

Herein lies another problem with using extremely low estimates of reliability in correction equations—the gap between observed correlations (reality) and corrected correlations (theory) is simply too large to be reasonably bridged in practice. The corrected coefficient tells us what to expect if we had an infinite number (or a very large number) of (psychometrically parallel) supervisors available to evaluate each employee. Is such an assumption reasonable? Isn't the average employee typically rated by only one or two supervisors (at the most)?

Stated another way, does it make sense to anchor our science on abjectly poor measurement, then make "corrections" for that poor measurement, and finally draw inferences as to what would happen if suddenly we were no longer subject to poor measurement? As Thouless (1939) noted 75 years ago when critiquing the use of "corrected" coefficients, "A prophet must be judged by his success in foretelling the future under the actual conditions in which he has to prophesy, not by an estimate of how well he would foretell the future if these conditions were improved" (p. 392).

In addition, by correcting for measurement error in only the criterion, organizational psychologists have placed a disproportional/asymmetric emphasis on the predictor while relegating the criterion to the status of second-class citizen in the nomological network of applied psychology. It

https://doi.org/10.1017/S1754942600006775 Published online by Cambridge University Press

serves to absolve psychologists from seeking or developing reliable criteria; and thus, it absolves psychologists from working to improve criterion theory and criterion measurement. Simply stated, this approach to test validation serves to exacerbate the long lamented criterion problem (Austin & Villanova, 1992; Binning & Barrett, 1989; Binning & LeBreton, 2009). It places a scientifically and psychometrically unhealthy emphasis on the predictor (i.e., we do not want to penalize the validity of the predictor due to poor measurement in the criterion). Implicit in this approach is a philosophy that predictors are more important for advancing our science (and our practice) than are criteria, that we should focus on improving our predictors if we want to improve our science and our practice. This approach to validation is not consistent with modern conceptualizations of theory building and validation. Indeed, for decades, our field has acknowledged (in principle) that accumulating the psychometric and construct validity evidence of our criteria is every bit as important as accumulating this evidence for our predictors (Binning & Barrett, 1989; Binning & LeBreton, 2009; Borman, 1991; James, 1973; SIOP, 2003). The time has come for our field to stop hiding behind asymmetric "corrections" and to start walking the scientific talk.

We also encourage researchers invoking correction procedures to use care when labeling the corrected or partially corrected coefficients. Coefficients that have been corrected should be so denoted ($\hat{\rho}$) rather than simply labeled as observed correlation coefficients (r) or referred to as "validities" without clearly articulating the various corrections made to these correlations (cf. Hunter & Hunter, 1984; Table 10; Schmidt & Hunter, 1998, Table 1). Labeling corrected coefficients as uncorrected coefficients could lead some psychologists (or HR managers) to draw improper inferences from meta-analyses. Such improper inferences could have serious consequences for I–O psychology, not only in terms of future theory construction but also in terms of practical policy decisions affecting I–O psychologists working in organizations (see for example the discussion by Professor Neal Schmitt involving the use of corrected vs. uncorrected correlations; p. 694, Morgeson et al., 2007).

Finally, we would like to remind the reader that it is only appropriate to interpret "corrected" correlations when (a) accurate estimates of reliability are available and (b) the causes or sources of variability in the reliability estimates are identical to the causes or sources of variability in the observed correlations (cf. Cronbach, Gleser, Nanda, & Rajaratnam, 1972; DeShon, 1998, 2003; Lord & Novick, 1968; Murphy & DeShon, 2000a; Winne & Belfry, 1982).

To summarize, corrections implemented via VG or SEM provide researchers with a powerful tool for glimpsing possible relationships among latent constructs. However, these corrections are not a panacea for applied psychologists. VG and SEM are statistical tools, just like *t*-tests, factor analysis, or social network analysis. They can be used or misused. And, like any statistical procedure they are no better than the data they analyze. Garbage in = garbage out.

Multivariate Validation

For the psychologist seeking to draw C-plane inferences, we would add that one of the advantages of latent variable SEM is the ability of the researcher to simultaneously model relationships among multiple predictors and multiple criteria. The days of validity studies using "overall job performance" are quickly becoming a thing of the past (cf. Aguinis, 2013; Borman, 1991; LeBreton & Tonidandel, 2008; Murphy & Shiarella, 1997; Rotundo & Sackett, 2002). SEM represents one of the more promising techniques available to psychologists seeking to answer calls for expanding validation work into fully multivariate frameworks (Murphy & Shiarella, 1997). However, the problems noted above with a lack of concordance in variance components will be exacerbated when moving to a multivariate

https://doi.org/10.1017/S1754942600006775 Published online by Cambridge University Press

https://doi.org/10.1017/S1754942600006775 Published online by Cambridge University Press

Table 5. *Return to Psychometrics Standards: On the Proper and Improper Estimation and Interpretation of "Corrected" Correlations*

Source	Observations and recommendations
Spearman (1910)	"This elimination [of errors from the observed correlation] is needed . . . in sciences such as psychology and sociology . . . But it is just in such sciences that the assumptions [necessary to calculate the disattenuated relationship] become most dangerous." (p. 273) "In view of the easy statistical elimination of the accidental errors, it might be thought no longer necessary to make long careful measurements. But this would be a grave mistake; for as seen from [the probable error of sampling equation], such accidents swell the correct coefficients probable error. Hence, the function of the formula here proposed is by no means to replace accurateness of original data, but on the contrary to emphasize the necessity of such accurateness, to estimate the degree of its realization, and only in the last instance to supplement its defectiveness." (p. 277)
Lanier (1927)	"While [the correction for attenuation] is sound, the assumptions that underlie it should be borne in mind and the all too prevalent practice of correcting for attenuation a mere chance coefficient of correlation should be discouraged." (p. 74) "In psychological and educational research there is at the present time a distinct tendency towards substituting mathematical treatment of data for careful experimentation and control of conditions." (p. 110) "Trained statisticians usually are aware of the limitations relating to the use of their methods . . . but the untrained individual who has merely the knowledge of how to manipulate the devices placed in his hands is in danger of straying, and the untrained person is not alone in this respect . . . When one reads some of the experimentation of Spearman, and notes how he corrects for attenuation coefficients of correlation from an exceeding small number of a cases and proceeds to far-reaching generalizations from the results, one cannot resist the notion that somehow the perfection existing in his hypothetical mathematical world has blinded him to the limitations of his experimental data." (p. 110)
Thouless (1939)	"It is, for example, a common problem in educational psychology to discover how successfully the results of school entrance examinations predict subsequent school success. Here the answer to our problem is given by the correlation coefficient uncorrected for attenuation and the corrected coefficient would be definitely wrong . . . A prophet must be judged by his success if foretelling the future under the actual conditions in which he has to prophesy, not by an estimate of how well he would foretell the future if these conditions were improved." (p. 392) "Whenever we want to determine the amount of correlation between test measurements (and not the amount of correlation between the abilities or other mental functions measured by the tests), then the attenuated coefficient must be used." (p. 392) "Correction for attenuation is wrong if what is required is the correlation between two test scores . . . It is necessary only when we want to know the absolute value of the correlation coefficient between the two tested [mental functions or constructs]. Consideration of the meaning of the correlation coefficient suggests that this knowledge is probably less often of importance than is commonly supposed." (p. 401)

https://doi.org/10.1017/S1754942600006775 Published online by Cambridge University Press

Table 5. *Continued*

Source	Observations and recommendations
Gulliksen (1950)	"We should note carefully just what we are doing when using this [disattenuation] equation. It is an estimate of the correlation between test and criterion if both could be made perfectly reliable by *lengthening the test and the criterion measure indefinitely*." (p. 102; italics added)
McNemar (1962)	"Correlation coefficients corrected for attenuation are of theoretical importance in the analysis of relationships in that allowance can be made for variable errors in measurement, but such corrected *r*s are of *little practical value* since they cannot be used in prediction equations. The prediction of one variable from another and the accompanying error of estimate must necessarily be based on obtained, or fallible, rather than true scores." (p. 153; italics added)
	"Corrected *r*s greatly in excess of unity have been reported. Such absurd results lead us to ask whether the assumptions [for disattenuation] have been met, but this question should be raised concerning any corrected *r*, even though it does not exceed unity, since the assumptions are difficult to meet." (p. 154)
Lord and Novick (1968)	". . . the so-called attenuation formulas, . . . give the correlation between measurements "corrected" for the unreliability of one or both measurements . . . It would seem that a simple way to obtain the disattenuated correlation ρ_{TXTY} [the correlation between true scores] is to estimate the quantities ρ_{XY}, $\rho_{XX'}$, and $\rho_{YY'}$ . . . and then use [the standard equation for disattenuation]. Unfortunately this procedure is often unsatisfactory. Indeed, in some cases this approach has led to the impossible conclusion that $\rho_{TXTY} > 1$." (pp. 137–138)
	". . . the correction for attenuation tends to be an *overcorrection of unknown extent*. If we are to use the correction for attenuation precisely, we must use it only in conjunction with an experimental design that assures that essentially no more or less error variation is introduced into the estimate of $\rho_{XX'}$, and $\rho_{YY'}$ than is introduced into the estimate of ρ_{XY}." (p. 138; italics added)
Womer (1968)	"It is possible statistically to "correct for" the unreliability of a criterion . . . when one is faced with making inferences about behavior in the real world, *it is not particularly useful to know how predictive a test would be if criterion measures were perfect*." (p. 65; italics added)
Nunnally (1978)	"One could argue that the correction for attenuation provides a way of fooling oneself into believing that a better correlation has been found than that actually evidenced in the available data. (p. 237)
Allen and Yen (1979)	"The corrected values are the maximum coefficients that are obtained if measurement errors could be totally eliminated. Since total elimination of measurement error generally is not possible, we cannot expect to observe the values of correlations corrected for attenuation. Also, it is important that accurate reliability estimates be used in the correction for attenuation. *If a reliability is underestimated, a misleading overestimate of* ρ_{TXTY} *will be obtained*." (p. 99; italics added)

https://doi.org/10.1017/S1754942600006775 Published online by Cambridge University Press

Table 5. *Continued*

Source	Observations and recommendations
Ghiselli, Campbell, and Zedeck (1981)	"Although it's a *dangerous question* in some respects, we might legitimately ask what the correlation between two variables might be if one or both of their reliabilities could be improved. If we push this question *to the greatest extreme*, we would ask what the correlation would be between two variables if one or both of their reliabilities were perfect." (p. 239; both sets of italics added) "In effect, what both [the classical theory of true and error scores & domain sampling] models argue is that to measure a variable with perfect reliability, we should allow the measure to grow to an *infinite length*." (p. 239; italics added)
Winne and Belfry (1982)	"underestimates of reliability produce overestimated correlations corrected for attenuation, thereby making interpretations too liberal." (p. 126)
Rosenthal (1984) as quoted in DeShon (2003)	"correcting for artifacts is not the proper goal of meta-analysis. The purpose of meta-analysis is to teach us what is, not what might be some day in the best of all possible worlds when all of our variables might be perfectly measured." (p. 386)
Seymour (1988)	"Hydraulics functions [corrections for attenuation] are mathematical operations which have the effect of increasing the size and significance of any observed correlation . . . They are so called because they serve the same function as any hydraulic system: they push to the sky things which properly belong on the ground. The functions are useful for limited purposes, but it abuses those purposes to apply these 'corrections' without first satisfying the statistical conditions for their use." (p. 351) "Such serial 'corrections' confuse the essential distinction between what *might* be and what *is*. The observance of that distinction is the primary factor separating science from mere supposition. The forgetting of that distinction is the hallmark of validity generalization." (p. 352; italics in original source)
Muchinsky (1996)	"Spearman's . . . correction for attenuation was used in the same spirit as the Spearman-Brown prophesy formula, which was developed to estimate what effect adding more items to a test would have on the test's reliability. Both formulae served as diagnostic aids to assist the researcher in test construction." (p. 66)
Schmitt and Chan (1998)	"this estimate of the "true" relationship [furnished by the disattenuated correlation] is very important when one is interested in *scientific questions* related to the actual *interrelationships among constructs*." (p. 144; italics added) "Corrected validities are very important in *theoretical discussions* because they represent the true underlying relationships between *constructs*." (p. 145; italics added)
McDonald (1999)	"Spearman's own example illustrates a *danger*: Underestimates of reliability coefficients may *overcorrect*, and give a projected gain in correlation that is *quite unattainable by any practical attempts* to eliminate error." (p. 134; italics added; this comment was made when discussing one of Spearman's [1904] early examples, wherein he obtained a corrected bivariate correlation of 1.03)

https://doi.org/10.1017/S1754942600006775 Published online by Cambridge University Press

Table 5. *Continued*

Source	Observations and recommendations
	"With care, [corrections for measurement error may be used] . . . A safe procedure would require the following three important conditions: 1. An unbiased and appropriate estimate of reliability should be used. . . 2. It is theoretically feasible to augment the sets of items [raters in the context of the current paper] with further items from the same conceptual domains. 3. Estimates of the reliability coefficients are obtained from the same sample that yields the correlation to be corrected, because such coefficients typically vary within the population." (p. 134)
DeShon (2003)	"The correlation between true scores cannot be greater than 1.0, which means that any disattenuated correlation that is greater than 1.0 is simply wrong—it is an inaccurate, overestimate of the correlation between true scores. If it occurs, the only reasonable interpretation is that the estimate of error used to disattenuate the correlation coefficient is an overestimate of the magnitude of measurement error (i.e., *the reliability coefficient is too small*)." (p. 384; italics added)
Oswald and McCloy (2003)	"Assessing the relationships among constructs more accurately requires developing better measures of those constructs—measures that are less contaminated and less deficient—rather than relying on statistical corrections." (p. 320)
Murphy and Davidshofer (2005)	"although theoretically justified, the correction for attenuation is often criticized on practical grounds. In practice, the correction almost always overestimates the actual correlation between *x* and *y* . . . In addition, *the correction formula assumes the impossible: that one can develop tests with no measurement error*." (pp. 137–138; italics added)

latent variable framework (DeShon, 1998). Thus, SEM is not a panacea for estimating C-plane relationships. Instead, like meta-analysis, SEM will only be as useful, informative, and accurate as the data being analyzed.

Conclusion

By using a criterion reliability of .52, the boost in validity afforded to VG studies involving a single predictor renders a nonsensical result to a VG analysis involving multiple predictors. It is reasonable to conclude that past VG studies, which relied on such a dubious criterion reliability value, have a less than secure statistical foundation. These studies should be revisited and their conclusions reexamined from the standpoint that the use of an unacceptably low reliability of .52 in corrections for attenuation, likely produced grossly inflated validities. This would be a first step in getting back into the judgment process wherein we again apply reasonable standards to criterion reliabilities.

A second step would be for applied psychologists to revisit the counsel proffered by statisticians, psychometricians, and perhaps even the occasional litigator concerning how to interpret correlations "corrected" for attenuation and under what circumstances it is even appropriate to make such corrections (see Table 5).

Finally, what might be the possible outcomes for our field if we all agreed to adhere to the standards for measurement (and interpretation) presented in Tables 1 and 5? We see at least three possible outcomes:

1. Supervisory ratings might disappear as criteria used in validation research, in the same way that projective tests have largely disappeared from applied psychology, or
2. Psychologists might take seriously the problems with ratings and work relentlessly to improve the psychometric characteristics of supervisory ratings. Given 100 years of research documenting "criterion problems" (Austin & Villanova, 1992) and over 60 years of problems with ratings (Guilford, 1954), we would be delightfully surprised if this outcome came to pass, or
3. Psychologists might take seriously the problems with ratings and question whether interrater correlations between two supervisors provide the most accurate and reasonable estimates of reliability. We suspect that researchers would conclude that .52 is not a credible estimate and thus work to find more appropriate ways to estimate the reliability of supervisory ratings (cf. DeShon, 2003; LeBreton et al., 2003; Murphy & DeShon, 2000a; Putka et al., 2008).

Future VG studies should invoke reasonable psychometric standards for measures of both predictors and criteria, as should studies employing latent variable procedures such as LISREL. We need to stop basing scientific conclusions on large corrections engendered by abjectly poor measurement, and we need to stop misinterpreting corrected (or even partially corrected) correlations.

References

Aguinis, H. A. (2013). *Performance management* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Aiken, L. R. (1988). *Psychological testing and assessment* (6th ed.). Boston, MA: Allyn and Bacon.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Prospect Heights, IL: Waveland Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Anastasi, A. (1968). *Psychological testing* (3rd ed.). New York, NY: The MacMillan Company.

Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917-1992. *Journal of Applied Psychology*, *77*, 836–874.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*, 1–25.

Bhaskar-Shrinivas, P., Harrison, D. A., Shaffer, M., & Luk, D. M. (2005). Input-based and time-based models of international adjustment: Meta-analytic evidence and theoretical extensions. *Academy of Management Journal*, *48*, 257–281.

https://doi.org/10.1017/S1754942600006775 Published online by Cambridge University Press

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology*, *74*, 478–494.

Binning, J. F., & LeBreton, J. M. (2009). Coherent conceptualization is useful for many things, and understanding validity is one of them. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *2*, 486–492.

Borman, W. C. (1991). Job behavior, performance, and effectiveness. In M. D. Dunnette, & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol 2, pp. 271–326). Palo Alto, CA: Consulting Psychologists Press.

Bowling, N. A., Eschleman, K. J., Wang, Q., Kirkendall, C., & Alarcon, G. (2010). A meta-analysis of the predictors and consequences of organization-based self-esteem. *Journal of Occupational and Organizational Psychology*, *83*, 601–626.

Bradshaw, F. F. (1930). The American Council on Education rating scale. *Archives of Psychology*, *199*, 80.

Campion, M. A., Campion, J. E., & Hudson, J. P., Jr. (1994). Structured interviewing: A note on incremental validity and alternative question types. *Journal of Applied Psychology*, *79*, 998–1002.

Cardy, R. L., & Dobbins, G. H. (1994). *Performance appraisal: Alternative perspectives*. Cincinnati, OH: South-Western.

Cascio, W. F., & Aguinis, H. (2005). *Applied psychology in human resource management* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Chiaburu, D. S., & Harrison, D. A. (2008). Do peers make the place? Conceptual synthesis and meta-analysis of coworker effects on perceptions, attitudes, OCBs, and performance. *Journal of Applied Psychology*, *93*, 1082–1103.

Colquitt, J. A., Conlon, D. E., Wesson, M. J., Porter, C. O. L. H., & Ng, K. Y. (2001). Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. *Journal of Applied Psychology*, *86*, 425–445.

Colquitt, J. A., Scott, B. A., & LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. *Journal of Applied Psychology*, *92*, 909–927.

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin*, *90*, 218–244.

Cooper-Hakim, A., & Viswesvaran, C. (2005). The construct of work commitment: Testing an integrative framework. *Psychological Bulletin*, *131*, 241–259.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: John Wiley & Sons.

DeShon, R. P. (1998). A cautionary note on measurement error corrections in structural equation models. *Psychological Measurement*, *3*, 412–423.

DeShon, R. P. (2003). A generalizability theory perspective on measurement error corrections in validity generalization. In K. R. Murphy (Ed.), *Validity generalization: A critical review* (pp. 365–402). Mahwah, NJ: Lawrence Erlbaum.

Erdogan, B., Bauer, T. N., Truxillo, D. M., & Mansfield, L. R. (2012). Whistle while you work: A review of the life satisfaction literature. *Journal of Management*, *38*, 1038–1083.

Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology*, *66*, 127–148.

Ford, M. T., Cerasoli, C. P., Higgins, J. A., & Decesare, A. L. (2011). Relationships between psychological, physical, and behavioural health and work performance: A review and meta-analysis. *Work & Stress*, *25*, 185–204.

Fried, Y. (1991). Meta-analytic comparison of the job diagnostic survey and job characteristics inventory as correlates of work satisfaction and performance. *Journal of Applied Psychology*, *76*, 690–697.

Fried, Y., Shirom, A., Gilboa, S., & Cooper, C. L. (2008). The mediating effects of job satisfaction and propensity to leave on role stress-job performance relationships: Combining meta-analysis and structural equation modeling. *International Journal of Stress Management*, *15*, 305–328.

Ghiselli, E. E., & Brown, C. W. (1955). *Personnel and industrial psychology* (2nd ed.). New York, NY: McGraw-Hill.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco, CA: W. H. Freeman & Company.

Guilford, J. P. (1936). *Psychometric methods*. New York, NY: McGraw-Hill.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York, NY: McGraw-Hill.

Guilford, J. P., & Fruchter, B. (1973). *Fundamental statistics in education and psychology* (5th ed.). New York, NY: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: John Wiley & Sons.

Hermelin, E., & Robertson, I. T. (2001). A critique and standardization of meta-analytic validity coefficients in personnel selection. *Journal of Occupational and Organizational Psychology*, *74*, 253–277.

Hoobler, J. M., Hu, J., & Wilson, M. (2010). Do workers who experience conflict between the work and family domains hit a "glass ceiling"?: A meta-analytic investigation. *Journal of Vocational Behavior*, *77*, 481–494.

Huffcutt, A. I., & Arthur, W., Jr. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology*, *79*, 184–190.

Huffcutt, A. I., Roth, P. L., & McDaniel, M. A. (1996). A meta-analytic investigation of cognitive ability in employment interview evaluations: Moderating characteristics and implications for incremental validity. *Journal of Applied Psychology*, *81*, 459–473.

Hunter, J. E. (1983). *Test validation for 12,000 jobs: An application of job classification and validity generalization analysis to the general aptitude test battery*. USES test research report no. 45. Division of Counseling and Test Development Employment and Training Administration. Washington, DC: U.S. Department of Labor.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, *96*, 72–98.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage.

https://doi.org/10.1017/S1754942600006775 Published online by Cambridge University Press

James, L. R. (1973). Criterion models and construct validity for criteria. *Psychological Bulletin*, *80*, 75–83.
James, L. R., & LeBreton, J. M. (2012). *Assessing the implicit personality through conditional reasoning*. Washington, DC: American Psychological Association.
James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal analysis: Assumptions, models, and data*. Beverly Hills, CA: Sage.
Joseph, D. L., & Newman, D. A. (2010). Emotional intelligence: An integrative meta-analysis and cascading model. *Journal of Applied Psychology*, *95*, 54–78.
Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods*, *9*, 202–220.
Landy, F. J. (1985). *Psychology of work behavior* (3rd ed.). Homewood, IL: Dorsey Press.
Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist*, *41*, 1183–1192.
Landy, F. J., & Farr, J. L. (1980). Performance ratings. *Psychological Bulletin*, *87*, 72–107.
Lanier, L. H. (1927). Prediction of the reliability of mental tests and tests of special abilities. *Journal of Experimental Psychology*, *10*, 69–113.
Latham, G. P., & Wexley, K. N. (1981). *Increasing productivity through performance appraisal*. Reading, MA: Addison-Wesley.
LeBreton, J. M., Burgess, J. R. D., Kaiser, R. B., Atchley, E. K. P., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods*, *6*, 80–128.
LeBreton, J. M., & Senter, J. L. (2008). Answers to twenty questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*, 815–852.
LeBreton, J. M., & Tonidandel, S. (2008). Multivariate relative importance: Extending relative weight analysis to multivariate criterion spaces. *Journal of Applied Psychology*, *93*, 329–345.
Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest*, *1*, 27–66.
Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
Margenau, H. (1950). *The nature of physical reality*. New York, NY: McGraw Hill.
McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.
McNemar, Q. (1962). *Psychological statistics* (3rd ed.). New York, NY: John Wiley & Sons.
Meriac, J. P., Hoffman, B. J., Woehr, D. J., & Fleisher, M. S. (2008). Further evidence for the validity of assessment center dimensions: A meta-analysis of the incremental criterion-related validity of dimension ratings. *Journal of Applied Psychology*, *93*, 1042–1052.
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749.
Michel, J. S., Mitchelson, J. K., Kotrba, L. M., LeBreton, J. M., & Baltes, B. B. (2009). A comparative test of work-family conflict models and critical examination of work-family linkages. *Journal of Vocational Behavior*, *74*, 199–218.
Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology*, *60*, 683–729.
Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement*, *56*, 63–75.
Murphy, K. R., & Balzer, W. K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology*, *74*, 619–624.
Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Thousand Oaks, CA: Sage.
Murphy, K. R., & Davidshofer, C. O. (2005). Psychological testing: Principles and applications (6th ed.). Upper Saddle River, NJ: Prentice Hall.
Murphy, K. R., & DeShon, R. (2000a). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology*, *53*, 873–900.
Murphy, K. R., & DeShon, R. (2000b). Progress in psychometrics: Can industrial and organizational psychology catch up? *Personnel Psychology*, *53*, 913–924.
Murphy, K. R., & Shiarella, A. H. (1997). Implications of the multidimensional nature of job performance for the validity of selection tests: Multivariate frameworks for studying test validity. *Personnel Psychology*, *50*, 823–854.
Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.
Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.
Oh, I., Wang, G., & Mount, M. K. (2011). Validity of observer ratings of the five-factor model of personality traits: A meta-analysis. *Journal of Applied Psychology*, *96*, 762–773.
Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology*, *78*, 679–703.
Oswald, F. L., & McCloy, R. A. (2003). Meta-analysis and the art of the average. In K. R. Murphy (Ed.), *Validity generalization: A critical review* (pp. 311–338). Mahwah, NJ: Lawrence Erlbaum.
Ployhart, R. E., Schneider, B., & Schmitt, N. (2006). *Staffing organizations: Contemporary practice and theory* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.
Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology*, *93*, 959–981.
Riggle, R. J., Edmondson, D. R., & Hansen, J. D. (2009). A meta-analysis of the relationship between perceived organizational support and job outcomes: 20 years of research. *Journal of Business Research*, *62*, 1027–1030.
Rockstuhl, T., Dulebohn, J. H., Ang, S., & Shore, L. M. (2012). Leader-member exchange (LMX) and

https://doi.org/10.1017/S1754942600006775 Published online by Cambridge University Press

culture: A meta-analysis of correlates of LMX across 23 countries. *Journal of Applied Psychology*, *97*, 1097–1130.

Rosenthal, R. (1984). *Meta-analysis procedures for social research*. Beverly Hills, CA: Sage.

Rothstein, H. R., Schmidt, F. L., Erwin, F. W., Owens, W. A., & Sparks, C. P. (1990). Biographical data in employment selection: Can validities be made generalizable? *Journal of Applied Psychology*, *75*, 175–184.

Rotundo, M., & Sackett, P. R. (2002). The relative importance of task, citizenship, and counterproductive performance to global ratings of job performance. *Journal of Applied Psychology*, *87*, 66–80.

Rugg, H. O. (1921). Is the rating of human character practicable? *Journal of Educational Psychology*, *12*, 425–438.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*, 413–428.

Salas, E., Rozell, D., Mullen, B., & Driskell, J. E. (1999). The effect of team building on performance: An integration. *Small Group Research*, *30*, 309–329.

Schmidt, F. L. (1992). What do data really mean?: Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist*, *47*, 1173–1181.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, *62*, 529–540.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*, 262–274.

Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology*, *71*, 432–439.

Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology*, *53*, 901–912.

Schmitt, N., & Chan, D. (1998). *Personnel selection: A theoretical approach*. Thousand Oaks, CA: Sage.

Seymour, R. T. (1988). Why plaintiffs' counsel challenge tests, and how they can successfully challenge the theory of "validity generalization." *Journal of Vocational Behavior*, *33*, 331–364.

Smith, P. C. (1976). Behavior, results, and organizational effectiveness: The problem of the criteria. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 745–775). Chicago, IL: Rand-McNally College.

Society for Industrial and Organizational Psychology (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

Spearman, C. (1910). Correlation calculated with faulty data. *British Journal of Psychology*, *3*, 271–295.

Symonds, P. M. (1931). *Diagnosing personality and conduct*. New York, NY: D. Appleton-Century Company, Inc.

Thomas, J. P., Whitman, D. S., & Viswesvaran, C. (2010). Employee proactivity in organizations: A comparative meta-analysis of emergent proactive constructs. *Journal of Occupational and Organizational Psychology*, *83*, 275–300.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*, 25–29.

Thouless, R. H. (1939). The effects of errors of measurement on correlation coefficients. *British Journal of Psychology: General Section*, *29*, 383–403.

Tornau, K., & Frese, M. (2013). Construct clean-up in proactivity research: A meta-analysis on the nomological net of work-related proactivity concepts and their incremental validities. *Applied Psychology: An International Review*, *62*, 44–96.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, *81*, 557–574.

Weiner, E. A., & Stewart, B. J. (1984). *Assessing individuals: Psychological and educational tests and measurements*. Boston, MA: Little, Brown, & Company.

Winne, P. H., & Belfry, M. J. (1982). Interpretive problems when correcting for attenuation. *Journal of Educational Measurement*, *19*, 125–134.

Womer, F. B. (1968). *Basic concepts in testing*. Boston, MA: Houghton Mifflin Company.

https://doi.org/10.1017/S1754942600006775 Published online by Cambridge University Press